



LHC Data Grid

The GriPhyN Perspective

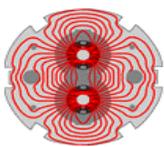
GriPhyN



Data Intensive Science

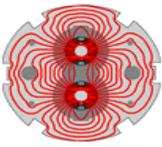
Paul Avery
University of Florida
<http://www.phys.ufl.edu/~avery/>
avery@phys.ufl.edu

**DOE/NSF Baseline Review of US-CMS
Software and Computing
Brookhaven National Lab
Nov. 15, 2000**





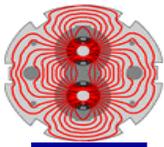
Fundamental IT Challenge



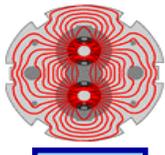
“Scientific communities of thousands, distributed globally, and served by networks with bandwidths varying by orders of magnitude, need to extract small signals from enormous backgrounds via computationally demanding (Teraflops-Petaflops) analysis of datasets that will grow by at least 3 orders of magnitude over the next decade, from the 100 Terabyte to the 100 Petabyte scale.”



GriPhyN = App. Science + CS + Grids



- ➔ **Several scientific disciplines**
 - ◆ **US-CMS** High Energy Physics
 - ◆ **US-ATLAS** High Energy Physics
 - ◆ **LIGO/LSC** Gravity wave research
 - ◆ **SDSS** Sloan Digital Sky Survey
- ➔ **Strong partnership with computer scientists**
- ➔ **Design and implement production-scale grids**
 - ◆ Maximize effectiveness of large, disparate resources
 - ◆ Develop common infrastructure, tools and services
 - ◆ Build on existing foundations: **PPDG project, Globus tools**
 - ◆ Integrate and extend capabilities of existing facilities
- ➔ **≈ \$70M total cost ⇒ NSF**
 - ◆ **\$12M: R&D**
 - ◆ **\$39M: Tier 2 center hardware, personnel**
 - ◆ **\$19M: Networking**

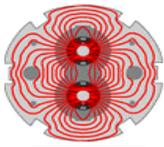


Importance of GriPhyN

- ➔ **Fundamentally alters conduct of scientific research**
 - ◆ **Old:** People, resources flow inward to labs
 - ◆ **New:** Resources, data flow outward to universities
- ➔ **Strengthens universities**
 - ◆ Couples universities to data intensive science
 - ◆ Couples universities to national & international labs
 - ◆ Brings front-line research to students
 - ◆ Exploits intellectual resources of formerly isolated schools
 - ◆ Opens new opportunities for minority and women researchers
- ➔ **Builds partnerships to drive new IT/science advances**
 - ◆ Physics ↔ Astronomy
 - ◆ Application Science ↔ Computer Science
 - ◆ Universities ↔ Laboratories
 - ◆ Fundamental sciences ↔ IT infrastructure
 - ◆ Research Community ↔ IT industry



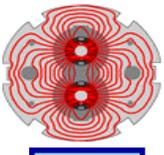
GriPhyN R&D Funded



- ➔ **NSF/ITR results announced Sep. 13**
 - ◆ \$11.9M from **Information Technology Research Program**
 - ◆ \$ 1.4M in matching from universities
 - ◆ Largest of all ITR awards
 - ◆ Joint NSF oversight from **CISE** and **MPS**
- ➔ **Scope of ITR funding**
 - ◆ Major costs for people, esp. students, postdocs
 - ◆ No hardware or professional staff for operations !
 - ◆ 2/3 CS + 1/3 application science
 - ◆ Industry partnerships being developed
 - **Microsoft, Intel, IBM, Sun, HP, SGI, Compaq, Cisco**



GriPhyN Institutions

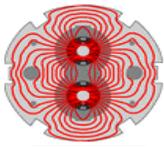


- ◆ **U Florida**
- ◆ **U Chicago**
- ◆ **Caltech**
- ◆ **U Wisconsin, Madison**
- ◆ **USC/ISI**
- ◆ **Harvard**
- ◆ **Indiana**
- ◆ **Johns Hopkins**
- ◆ **Northwestern**
- ◆ **Stanford**
- ◆ **Boston U**
- ◆ **U Illinois at Chicago**
- ◆ **U Penn**
- ◆ **U Texas, Brownsville**
- ◆ **U Wisconsin, Milwaukee**
- ◆ **UC Berkeley**

- ◆ **UC San Diego**
- ◆ **San Diego Supercomputer Center**
- ◆ **Lawrence Berkeley Lab**
- ◆ **Argonne**
- ◆ **Fermilab**
- ◆ **Brookhaven**

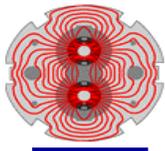


Funding for Tier 2 Centers



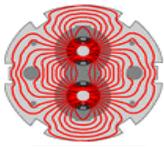
- GriPhyN/ITR has no funds for hardware
- NSF is proposal driven
 - ◆ How to get \$40M = hardware + software (all 4 expts)
 - ◆ Approx. \$24M of this needed for ATLAS + CMS
- Long-term funding possibilities not clear
 - ◆ Requesting funds for FY2001 for prototype centers
 - ◆ NSF ITR2001 competition (\$15M max)
 - ◆ Other sources?

Benefits of GriPhyN to LHC



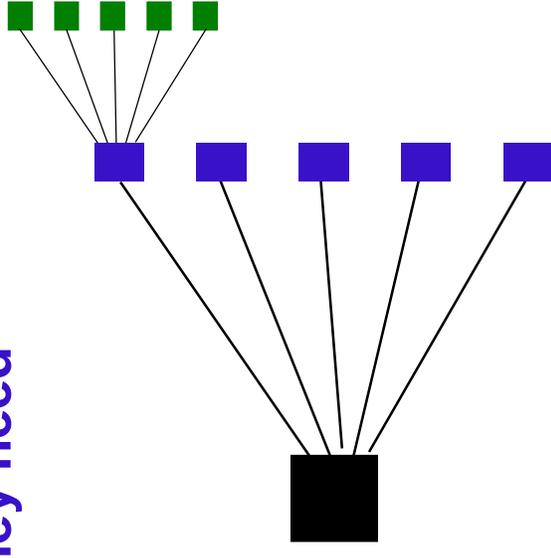
- ➔ **Virtual Data Toolkit for LHC computing**
- ➔ **Additional positions to support tool integration**
 - ◆ 1 postdoc + 1 student, Indiana (ATLAS)
 - ◆ 1 postdoc/staff, Caltech (CMS)
 - ◆ 2 scientists + 1 postdoc + 2 students, Florida (CMS)
- ➔ **Leverage vendor involvement for LHC computing**
 - ◆ Dense computing clusters (Compaq, SGI)
 - ◆ Load balancing (SGI, IBM)
 - ◆ High performance storage (Sun, IBM)
 - ◆ High speed file systems (SGI)
 - ◆ Enterprise-wide distributed computing (Sun)
 - ◆ Performance monitoring (SGI, IBM, Sun, Compaq)
 - ◆ Fault-tolerant clusters (SGI, IBM)
 - ◆ Network services (Cisco)

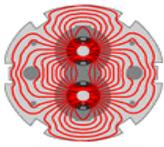
Why a Data Grid: Physical



- ➔ **Unified system: all computing resources part of grid**
 - ◆ Efficient resource use (manage scarcity)
 - ◆ Resource discovery / scheduling / coordination truly possible
 - ◆ “The whole is greater than the sum of its parts”
- ➔ **Optimal data distribution and proximity**
 - ◆ Labs are close to the (raw) data they need
 - ◆ Users are close to the (subset) data they need
 - ◆ Minimize bottlenecks

- ➔ **Efficient network use**
 - ◆ local > regional > national > oceanic
 - ◆ No choke points
- ➔ **Scalable growth**

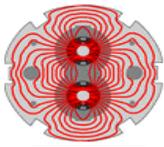




Why a Data Grid: Political

- ➔ **Central lab cannot manage / help 1000s of users**
 - ◆ Easier to leverage resources, maintain control, assert priorities at regional / local level
- ➔ **Cleanly separates functionality**
 - ◆ Different resource types in different Tiers
 - ◆ Organization vs. flexibility
 - ◆ Funding complementarity (NSF vs DOE), targeted initiatives
- ➔ **New IT resources can be added “naturally”**
 - ◆ Matching resources at Tier 2 universities
 - ◆ Larger institutes can join, bringing their own resources
 - ◆ Tap into new resources opened by IT “revolution”
- ➔ **Broaden community of scientists and students**
 - ◆ Training and education
 - ◆ Vitality of field depends on University / Lab partnership

GriPhyN Research Agenda

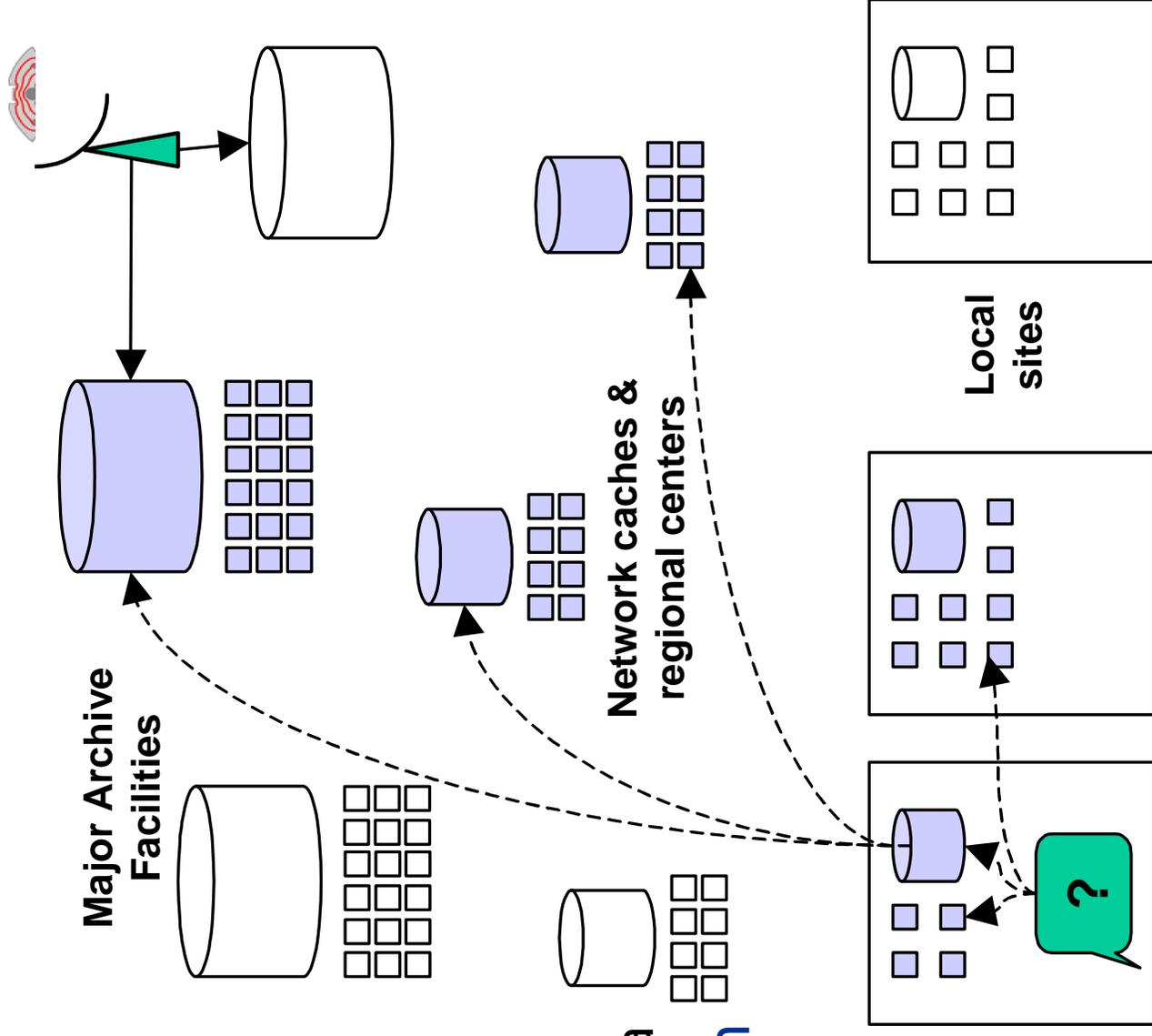


- ➔ **Virtual Data technologies**
 - ◆ Derived data, calculable via algorithm (e.g., 90% of HEP data)
 - ◆ Instantiated 0, 1, or many times
 - ◆ Fetch data vs. execute algorithm
 - ◆ Very complex (versions, consistency, cost calculation, etc)
- ➔ **Planning and scheduling**
 - ◆ User requirements (time vs cost)
 - ◆ Global and local policies + resource availability
 - ◆ Complexity of scheduling in dynamic environment (hierarchy)
 - ◆ Optimization and ordering of multiple scenarios
 - ◆ Requires simulation tools, e.g. MONARC

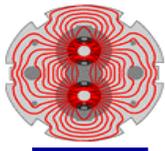


Virtual Data in Action

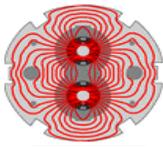
- Data request may
 - Compute locally
 - Compute remotely
 - Access local data
 - Access remote data
- Scheduling based on
 - Local policies
 - Global policies
- Local autonomy



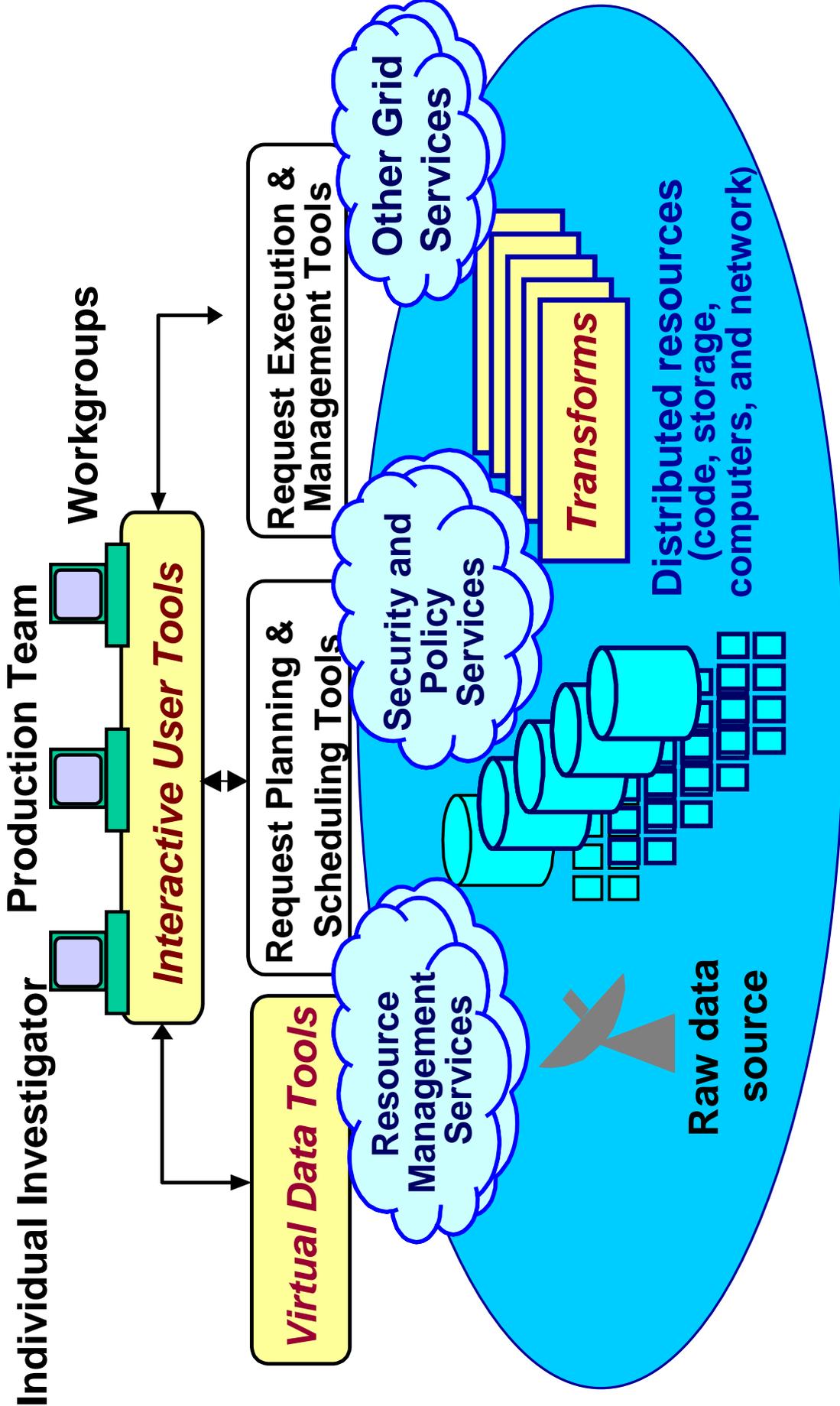
Research Agenda (cont.)

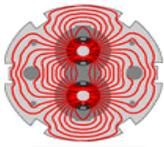


- ➔ **Execution management**
 - ◆ Co-allocation of resources (CPU, storage, network transfers)
 - ◆ Fault tolerance, error reporting
 - ◆ Agents (co-allocation, execution)
 - ◆ Reliable event service across Grid
 - ◆ Interaction, feedback to planning
- ➔ **Performance analysis**
 - ◆ Instrumentation and measurement of all grid components
 - ◆ Understand and optimize grid performance
- ➔ **Virtual Data Toolkit (VDT)**
 - ◆ VDT = virtual data services + virtual data tools
 - ◆ One of the primary deliverables of R&D effort
 - ◆ Ongoing activity + feedback from experiments (5 year plan)
 - ◆ Technology transfer mechanism to other scientific domains



PetaScale Virtual Data Grids

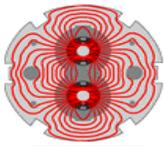




Tier 2 Hardware Costs

Year	2000	2001	2002	2003	2004	2005	2006
T2 #1	\$350K	\$90K	\$150K	\$120K	\$500K*	\$250K	\$250K
T2 #2		\$192K	\$320K	\$120K	\$500K*	\$250K	\$250K
T2 #3				\$350K	\$120K	\$500K*	\$250K
T2 #4						\$750K*	\$250K
T2 #5							\$750K*
Total	\$350K	\$282K	\$470K	\$590K	\$1200K	\$1750K	\$1750K

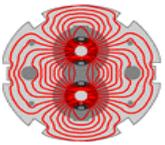
FY01-06 total: \$6.0M



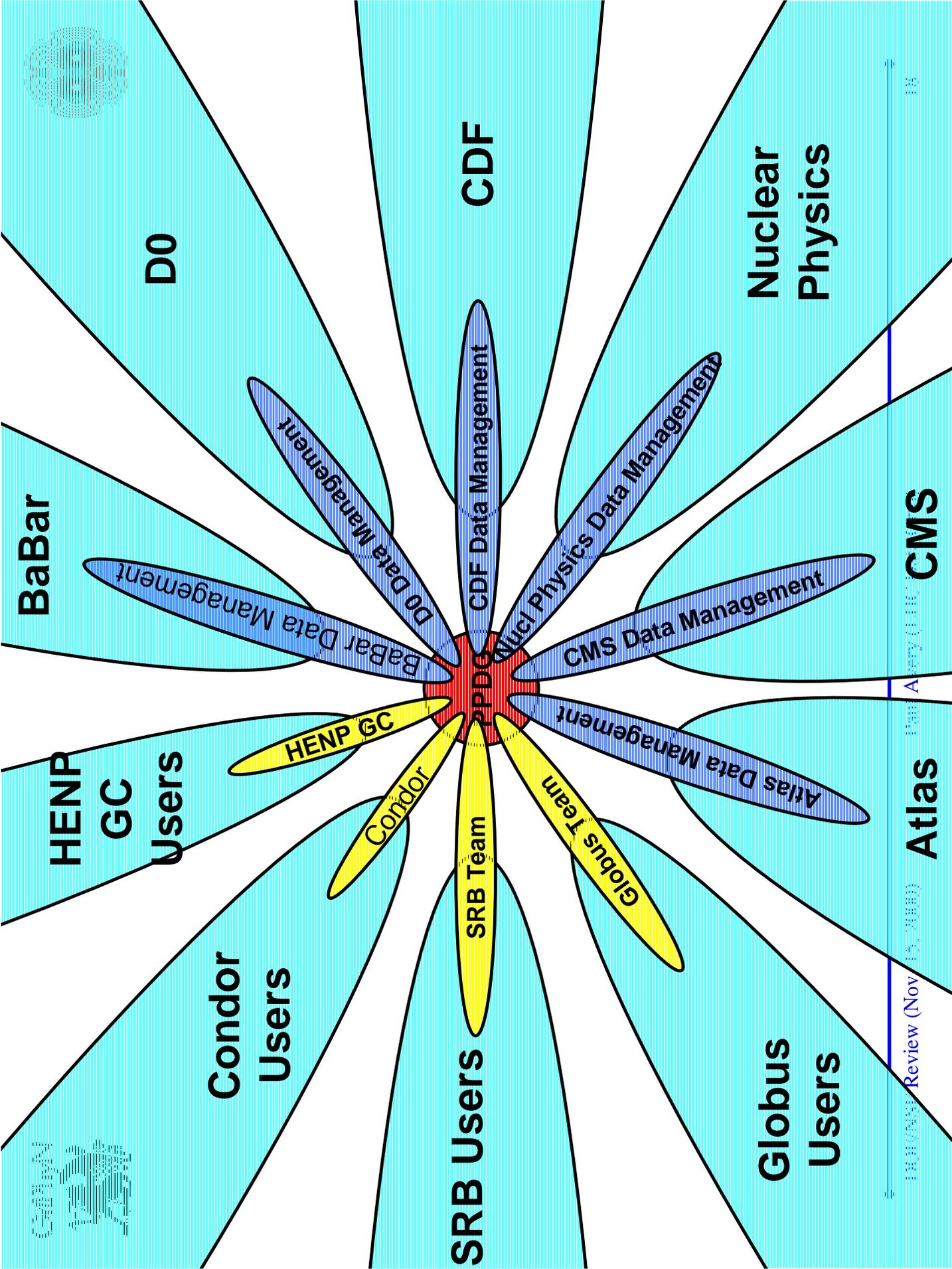
Tier 2 Personnel Costs

Year	2000	2001	2002	2003	2004	2005	2006
T2 #1	\$50K	\$200K	\$200K	\$200K	\$200K	\$200K	\$200K
T2 #2		\$200K	\$200K	\$200K	\$200K	\$200K	\$200K
T2 #3				\$200K	\$200K	\$200K	\$200K
T2 #4						\$200K	\$200K
T2 #5							\$200K
Total	\$50K	\$400K	\$400K	\$600K	\$600K	\$800K	\$1000K

FY01-06 total: \$3.8M

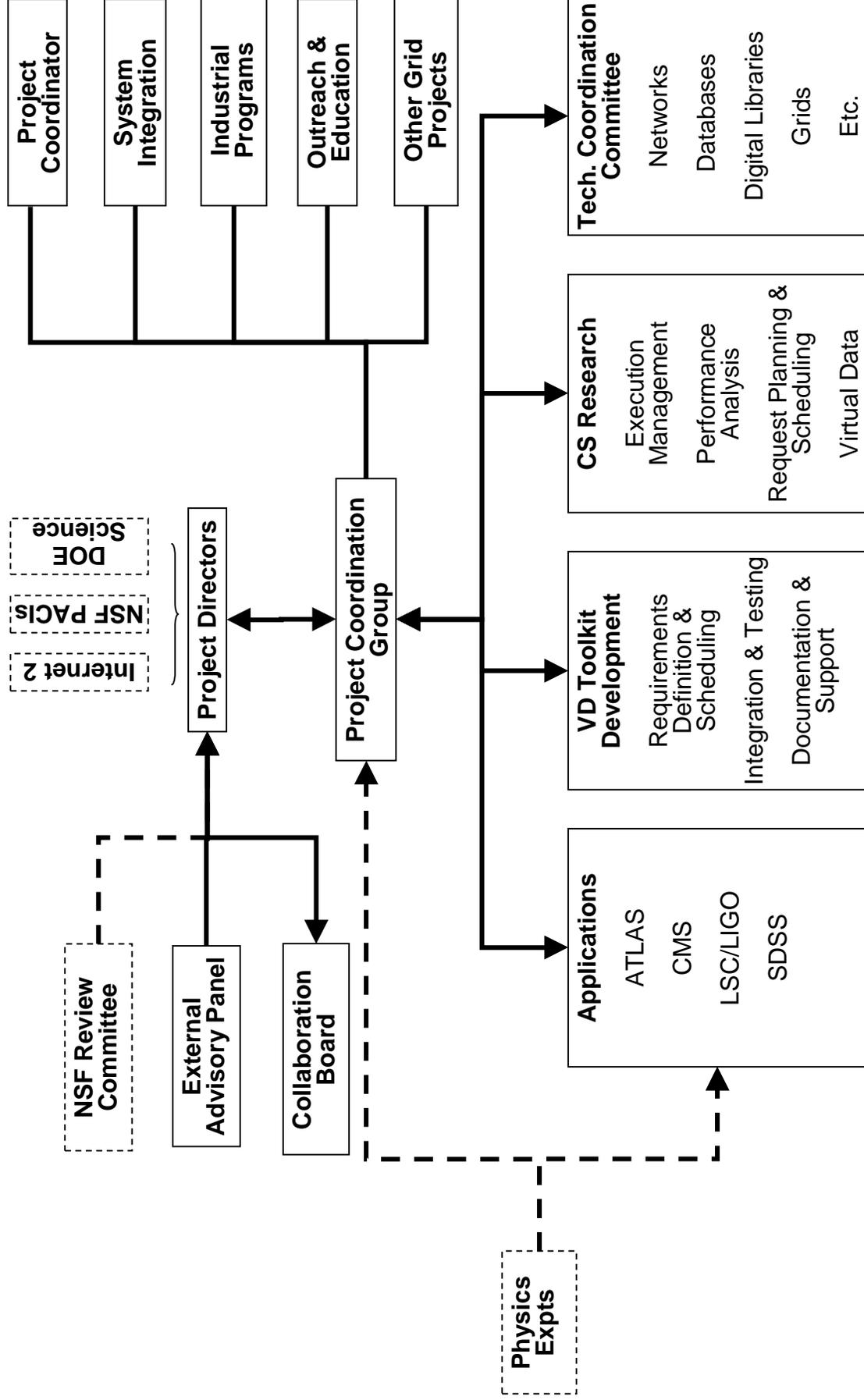
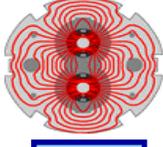


The Management Problem



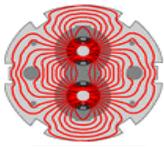


GriPhyN Org Chart





GriPhyN Management Plan

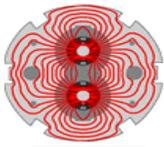


- ➔ **GriPhyN PMP incorporates CMS, ATLAS plans**
- ◆ Schedules, milestones
- ◆ Senior computing members in Project Coordination Group
- ◆ Formal liaison with experiments
- ◆ Formal presence in Application area

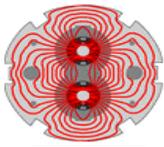
GriPhyN PMP:

“...the schedule, milestones and deliverables of the [GriPhyN] Project must correspond to those of the experiments to which they belong.”

Long-Term Grid Management



- ➔ **CMS Grid is a complex, distributed entity**
- ◆ **International scope**
- ◆ **Many sites from Tier 0 – Tier 3**
- ◆ **Thousands of boxes**
- ◆ **Regional, national, international networks**
- ➔ **Necessary to manage**
- ◆ **Performance monitoring**
- ◆ **Failures**
- ◆ **Bottleneck identification and resolution**
- ◆ **Optimization of resource use**
- ◆ **Change policies based on resources, CMS priorities**



Grid Management (cont)

- ➔ **Partition US-CMS grid management**
 - ◆ Can effectively separate USA Grid component
 - ◆ US Grid management coordinated with CMS
- ➔ **Tools**
 - ◆ Stochastic optimization with incomplete information
 - ◆ Visualization
 - ◆ Database performance
 - ◆ Many tools provided by GriPhyN performance monitoring
 - ◆ Expect substantial help from IT industry
- ➔ **Timescale and scope**
 - ◆ Proto-effort starting in 2004, no later than 2005
 - ◆ ≈4–6 experts needed to oversee US Grid
 - ◆ Expertise: ODBMS, networks, scheduling, grids, etc.